# APPENDIX A

# ISSUE PAPERS

# Issue Paper on Evaluating Representativeness of Exposure Factors Data

This paper is based on the Technical Memorandum dated March 4, 1998, submitted by Research Triangle Institute under U.S. EPA contract 68D40091.

## 1. INTRODUCTION

The purpose of this document is to discuss the concept of representativeness as it relates to assessing human exposures to environmental contaminants and to factors that affect exposures and that may be used in a risk assessment. (The factors, referred to as exposure factors, consist of measures like tapwater intake rates, or the amount of time that people spend in a given microenvironment.) This is an extremely broad topic, but the intent of this document is to provide a useful starting point for discussing this extremely important concept.

Section 2 furnishes some general definitions and notions of representativeness. Section 3 indicates a general framework for making inferences. Components of representativeness are presented in Section 4, along with some checklists of questions that can help in the evaluation of representativeness in the context of exposures and exposure factors. Section 5 presents some techniques that may be used to improve representativeness. Section 6 provides our summary and conclusions.

## 2. GENERAL DEFINITIONS/NOTIONS OF REPRESENTATIVENESS

Representativeness is defined in *American National Standard: Specifications and Guidelines for Quality Systems for Environmental Data and Environmental Technology Programs (ANSI/ASQC E4 - 1994)* as follows:

> The measure of the degree to which data accurately and precisely represent a characteristic of a population, parameter variations at a sampling point, a process condition, or an environmental condition.

Although Kendall and Buckland (*A Dictionary of Statistical Terms*, 1971) do not define representativeness, they do indicate that the term “representative sample” involves some confusion about whether this term refers to a sample “selected by some process which gives all samples an equal chance of appearing to represent the population” or to a sample that is “typical in respect of certain characteristics, however chosen.” Kruskal and Mosteller (1979) point out that representativeness does not have an unambiguous definition; in a series of three papers, they present and discuss various notions of representativeness in the scientific, statistical, and other literature, with the intent of clarifying the technical meaning of the term.

In Chapter 1 of the *Exposure Factors Handbook* (EFH), the considerations for including the particular source studies are enumerated and then these considerations are evaluated qualitatively at the end of each chapter (i.e., for each type of exposure factor data). One of the criteria is "representativeness of the population," although there are several other criteria that clearly relate to various aspects of representativeness. For example, these related criteria include the following:

| EFH Study Selection Criterion | EFH Perspective |
|---|---|
| focus on factor of interest | studies with this specific focus are preferred |
| data pertinent to U.S. | studies of U.S. residents are preferred |
| current information | recent studies are preferred, especially if changes over time are expected |
| adequacy of data collection period | generally the goal is to characterize long-term behavior |
| validity of approach | direct measurements are preferred |
| representativeness of the population | U.S. national studies are preferred |
| variability in the population | studies with adequate characterizations of variability are desirable |
| minimal (or defined) bias in study design | studies having designs with minimal bias are preferred (or with known direction of bias) |
| minimal (or defined) uncertainty in the data | large studies with high ratings on the above considerations are preferred |

## 3. A GENERAL FRAMEWORK FOR MAKING INFERENCES

Despite the lack of specificity of a definition of representativeness, it is clear in the present context that representativeness relates to the "comfort" with which one can draw inferences from some set(s) of extant data to the population of interest for which the assessment is to be conducted, and in particular, to certain characteristics of that population's exposure or exposure factor distribution. The following subsections provide some definitions of terms and attempt to break down the overall inference into some meaningful steps.

### 3.1 Inferences from a Sample to a Population

In this paper, the word **population** to refers to a set of units which may be defined in terms of person and/or space and/or time characteristics. The population can thus be defined in terms of its individuals' characteristics (defined by demographic and socioeconomic factors, human behavior, and study design) (e.g., all persons aged 16 and over), the spatial characteristics (e.g., living in Chicago) and/or the temporal characteristics (e.g., during 1997).

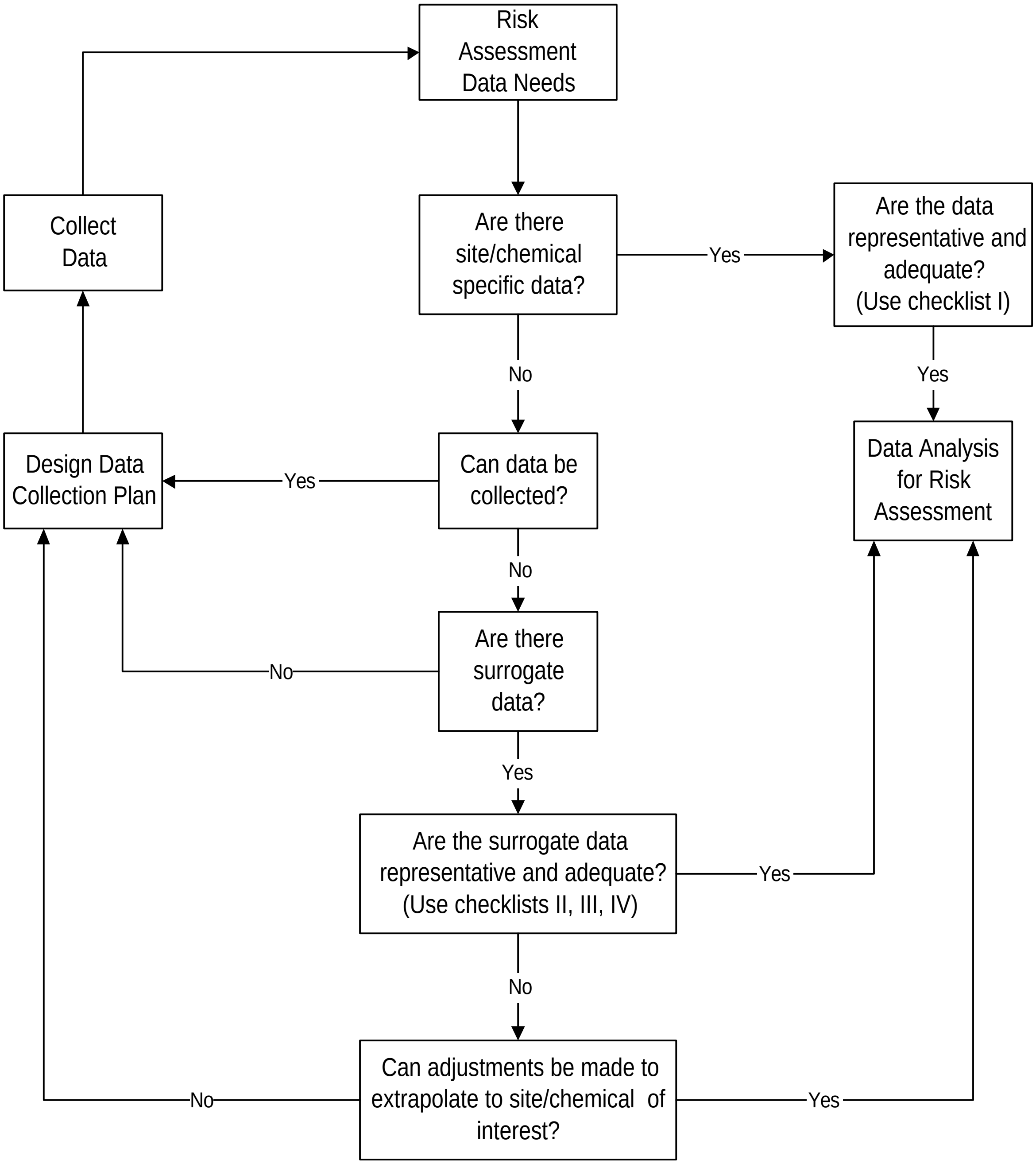
In conducting a risk assessment, the assessor needs to define the population of concern — that is, the set of units for which risks are to be assessed (e.g., lifetime risks of all U.S. residents). At a Superfund site, this population of concern is generally the population surrounding the site. In this document, the term population of concern refers to that population for which the assessor wishes to draw inferences. If it were practical, this is the population for which a census (a 100% sample) would exist or for which the assessor would conduct a probability-based study of exposures. Figure 1 provides a diagram of the exposure assessor decision process during the selection of data for an exposure assessment.

As depicted in figure 1, quite often it is not practical or feasible to obtain data on the population of concern and the assessor has to rely on the use of surrogate data. These data generally come from studies conducted by researchers for a variety of purposes. Therefore, the assessor's population of concern may differ from the surrogate population. Note that the population differences may be in any one (or more) of the characteristics described earlier. For example, the surrogate population may only cover a subset of the individuals in the assessor's population of concern (Chicago residents rather than U.S. residents). Similarly, the surrogate data may have been collected during a short period of time (e.g., days), while the assessor may be concern about chronic exposures (i.e., temporal characteristics).

The studies used to derive these surrogate data are generally designed with a population in mind. Since it may not be practical to sample everyone in that population, probability-based sampling are often conducted. This sampling scheme allows valid statistical (i.e., non-model-based) inferences, assuming there were no implementation difficulties (e.g., no nonresponse and valid measurements). Ideally, the implementation difficulties would not be severe (and hence ignored), so that these sampled individuals can be considered representative of the population. If there are implementation difficulties, adjustments are typically made (e.g., for nonresponse) to compensate for the population differences. Such remedies for overcoming inferential gaps are fairly well documented in the literature in the context of probability-based survey sampling (e.g., see Oh and Scheuren (1983)). If probability sampling is not employed, the relationships of the selected individuals for which data are sought and of the respondents for which data are actually acquired to the population for which the study was designed to address are unclear.

There are cases where probability-based sampling is used and the study design allows some model-based inferences. For instance, food consumption data are often obtained using surveys which ask respondents to recall food eaten over a period of few days. These data are usually collected throughout a one-year period to account for some seasonal variation in food consumption. Statistical inferences can then be made for the individuals surveyed within the time frame of study. For example, one can estimate the mean, the 90th percentile, etc. for the number

## Figure 1: Risk Assessment Data Collection Process


Risk Assessment Data Needs
Collect Data
Are there site/chemical specific data?
Yes
Are the data representative and adequate? (Use checklist I)
No
Yes
Design Data Collection Plan
Yes
Can data be collected?
Data Analysis for Risk Assessment
No
Are there surrogate data?
No
Yes
Are the surrogate data representative and adequate? (Use checklists II, III, IV)
Yes
No
Can adjustments be made to extrapolate to site/chemical of interest?
No
Yes

of days during which individuals were surveyed. However, if at least some of the selected individuals are surveyed multiple periods of time during that year, then a model-based strategy might allow estimation of a distribution of long-term (e.g., annual) consumption patterns.

If probability-based sampling is not used, model-based rather than statistical inferences are needed to extend the sample results to the population for which the study was designed.

In contrast to the inferences described above, which emanate from population differences and the sampling designs used in the study, there are two additional inferential aspects that relate to representativeness:

- The degree to which the study design is followed during its implementation
- The degree to which a measured value represents the true value for the measured unit

Both of these are components of measurement error. The first relates to an implementation error in which the unit selected for measurement is not precisely the one for which the measurement actually is made. For instance, the study's sampling design may call for people to record data for 24-hr periods starting at a given time of day, but there may be some departure from this ideal in the actual implementation. The second has to do with the inaccuracy in the measurement itself, such as recall difficulties for activities or imprecision in a personal air monitoring device.

## 4. COMPONENTS OF REPRESENTATIVENESS

As described above, the evaluation of how representative a data set is begins with a clear definition of the population of concern (the population of interest for the given assessment), with attention to all three fundamental characteristics of the population — individual, spacial, and temporal characteristics. Potential inferential gaps between the data set and the population of concern -- that is, potential sources of unrepresentativeness -- can then be partitioned both along these population characteristics. Components of representativeness are illustrated in Table1: the rows correspond to the inferential steps and the columns correspond to the population characteristics. The inferential steps are distinguished as being either internal or external to the source study.

### 4.1 Internal Components - Surrogate Data Versus the Study Population

After determining that a study provides information on the exposures or exposure factors of interest, it is important that the exposure assessor evaluate the representativeness of the surrogate study (or studies). This entails gaining an understanding of both the individuals sampled for the study and the degree to which the study achieved valid inferences to that population. The assessor should consider the questions in Checklist I in the appendix to help establish the degree of representativeness inherent to this internal component. In the context of the Exposure Factors Handbook (EFH), the representativeness issues listed in this checklist are presumably the types of considerations that led to selection of the source studies that appear in

## Table 1. Elements of Representativeness

| **Component of Inference** | **Population Characteristics** | | |
|---|---|---|---|
| | **Individual Characteristics** | **Spacial Characteristics** | **Temporal Characteristics** |
| **EXTERNAL TO STUDY** | | | |
| How well does the surrogate population represent the population of concern? | • Exclusion or limited coverage of certain segments of population of concern | • Exclusion or inadequate coverage of certain regions or types of areas (e.g., rural areas) that make up the population of concern | • Lack of currency<br>• Limited temporal coverage, including exclusion or inadequate coverage of seasons<br>• Inappropriate duration for observations (e.g., short-term measurements where concern is on chronic exposures) |
| **INTERNAL TO STUDY** | | | |
| How well do the individuals sampled represent the population of concern for the study? | • Imposed constraints that exclude certain segments of study population<br>• Frame inadequacy (e.g., due to lack of current frame information) | • Inadequate coverage (e.g., limited to single urban area) | • Limited temporal coverage (e.g., limited study duration)<br>• Inappropriate duration for observations |
| How well do the actual number of respondents represent the sampled population?<br>How well does the measured value represent the true value for the measured unit? | • Non-probability sample of persons<br>• Excessive nonresponse<br>• Inadequate sample size<br>• Behavior changes resulting from participation in study (Hawthorne effect)<br>• Measurement errors associated with people's ability/desire to respond accurately to questionnaire items<br>• Measurement error associated with within-specimen heterogeneity<br>• Inability to acquire physical specimen with exact size or shape or volume desired | • Non-probability sample of spatial units (e.g., convenience or judgmental siting of ambient monitors)<br>• Inaccurate identification of sampled location | • Non-probability sample of observation times<br>• Deviation in times selected vs. those measured or reported (e.g., due to schedule slippage, or incomplete response)<br>• Measurement errors related to time (e.g., recall difficulties for foods consumed or times in microenvironments) |

the EFH. As indicated previously, the focus for addressing representativeness in that context was national and long-term, which may or may not be consistent with the assessment of current interest.

### 4.2 External Components - Population of Concern Versus Surrogate Population

In many cases, the assessor will be faced with a situation in which the population of concern and surrogate population do not coincide in one or more aspects. To address this external factor of representativeness, the assessor needs to:

- determine the relationship between the two populations
- judge the importance of any discrepancies between the two populations
- assess whether adjustments can be made to reconcile or reduce differences.

To address these, the assessor needs to consider all characteristics of the populations. Relevant questions to consider are listed in Checklists II, III, and IV in the appendix for the individual, spacial, and temporal characteristics, respectively.

Each checklist contains several questions related to each of the above bullets. For example, the first few items of each checklist relate to the first item above (relationship of the two populations). There are several possible ways in which the two populations may relate to each other; these cases are listed below and can be addressed for each population dimension:

- Case 1: The population of concern and surrogate population are (essentially) the same
- Case 2: The population of concern is a subset of the surrogate population
  - Case 2a: The subset is a large and identifiable subset.
  - Case 2b: The subset is a small and/or unidentifiable subset.
- Case 3: The surrogate population is a subset of the population of concern.
- Case 4: The population of concern and surrogate population are disjoint.

Note that Case 2a implies that adequate data are available from the surrogate study to generate separate summary statistics (e.g., means, percentiles) for the population of concern. For example, if the population of concern was focused on children and the surrogate population was a census or large probability study of all U.S. residents, then children-specific summaries would be possible. In such a situation, Case 2a reverts to back to Case 1.

Case 2b will be typical of situations in which large-scale (e.g., national or regional) data are available but assessments are needed for local areas (or for acute exposures). As an example, suppose raw data from the National Food Consumption Survey (NFCS) can be used to form meaningful demographic subgroups and to estimate average tapwater consumption for such subgroups (e.g., see Section 5.1). If a risk assessment involving exposure from copper smelters is to be conducted for the southwestern U.S., for instance, tapwater consumption would probably be considered to be different for that area than for the U.S. as a whole, but the NFCS data for that area might be adequate.  If so, this would be considered Case 2a.  But if the risk assessment concerned workers at copper smelters, then an even greater discrepancy between the population of concern and the surrogate data might be expected, and the NFCS data would likely be regarded as inadequate, and more speculative estimates would be needed.

In contrast to Case 2, Case 3 will be typical of assessments that must use local and/or short-term data to extrapolate to regional or national scales and/or to long-term (chronic) exposures. Table 2 presents some hypothetical examples for each case.  Note that, as illustrated here and as implied by the bulleted items in Checklist IV, the temporal characteristics has two series of issues:  one that relates to the currency and the temporal coverage (study duration) of the source study relative to the population of concern time frame, and one that relates to the time unit of observation associated with the study.

Since most published references to the NFCS rely on the 1977-78 survey, exposure factor data based on that survey might well be considered as Case 4 with respect to temporal coverage, as trends such as consumption of bottled water and organic foods may not be well represented by 20 year-old data.  A possible approach in this situation would be to obtain data from several NFCSs, to compare or test for a difference between them, and to use them to extrapolate to the present or future.  The NFCS also illustrates the other temporal aspect — dealing with a time-unit mismatch of the data and the population of concern — since the survey involves three consecutive days for each person, while typically a longer-term estimate would be desired, e.g., a person-year estimate (e.g., see Section 5.2).

While determining the relationship of the two populations will generally be straightforward (first bullet), determining the importance of discrepancies and making adjustments (the second and third bullets) may be highly subjective and require an understanding of what factors contribute to heterogeneity in exposure factor values and speculation as to their influence on the exposure factor distribution.  Cases 1 and 2a are the easiest, of course.  In the other cases, it will generally be easier to speculate about how the mean and variability (perhaps expressed as a coefficient of variation (CV)) of the two populations may differ than to speculate on changes in a given percentile.  Considerations of unaffected portions of the population must also be factored into the risk assessor's speculation.  The difficulty in such speculation obviously increases dramatically when two or more factors affect heterogeneity, especially if the factors are anticipated to have opposite or dependent effects on the exposure factor values.  Regardless of how such speculation is ultimately reflected in the assessment (either through ignoring the population differences or by adjusting estimated parameters of the study population), recognition of the increased uncertainty should be incorporated into sensitivity analyses.  As a part of such an analysis, it would be instructive to determine risks, when, for each relevant factor (e.g., age category), several assessors independently speculate on the mean (e.g., a low, best guess, and high) and on the CV.

## Table 2.  Examples of Relationships Between the Population of Concern and the Surrogate Population

| Population Characteristics | Population | Case 1: Population of concern and surrogate population are the same | Case 2a: Population of concern is a subset of the surrogate population, and data on subset are available | Case 2b: Population of concern is a subset of surrogate population and data on subset are not available | Case 3: Surrogate population is a subset of the population of concern | Case 4: Population of concern and surrogate population are disjoint |
|---|---|---|---|---|---|---|
| Individual Characteristics | Population of concern: | U.S. residents | U.S. children | Asthmatic U.S. children | U.S. residents | U.S. children |
| | Surrogate population: | U.S. residents | U.S. residents + age data | U.S. residents | U.S. adults | U.S. adults |
| Spacial Characteristics | Population of concern: | U.S. | Northeast U.S. | Near hazardous waste sites | U.S. | U.S. |
| | Surrogate population: | U.S. | U.S. + region ID data | U.S. | Chicago | Netherlands |
| Temporal Characteristics and Currency | Population of concern: | one year, 1998 | summer, 1998 | 1998 days with smog | lifetime | future years |
| | Surrogate population: | one year, 1998 | one year, 1998 + season ID data | one year, 1998 | two summertime weeks, 1998 | 1996 |
| Temporal Observation Units | Population of concern: | person-days | eating occasions | eating occasions (acute) | lifetimes (chronic) | NA |
| | Surrogate population: | person-days | person-days + meal-specific data | person-3-day data | person-days | |

## 5. ATTEMPTING TO IMPROVE REPRESENTATIVENESS

### 5.1 Adjustments to Account for Differences in Population Characteristics or Coverage.

If there is some overlap in information available for the population of concern and the surrogate population (e.g., age distributions), then adjustments to the sample data can be made that attempt to reduce the bias that would result from directly applying the study results to the population of concern. Such methods of adjustment can all be generally characterized as "direct standardization" techniques, but the specific methodology to use depends on whether one has access to the raw data or only to summary statistics, as is often the case when using data from the Exposure Factors Handbook. With access to the raw data, the applicable techniques also depend on whether one wants to standardize to a single known population of concern distribution (e.g., age categories), to two or more marginal distributions known for the population of concern, or even to population of concern totals for continuous variables.

**Summary Statistics Available.** Suppose that the available data are summary statistics such as the mean, standard deviation, and various percentiles for an exposure factor of interest (e.g., daily consumption of tap water). Furthermore, suppose that these statistics are available for subgroups based on age, say age groups g = 1, 2, ..., G. Furthermore, suppose we know that the age distribution of the population of concern differs from that represented by the sample data. We can then estimate linear characteristics of the population of concern, such as the mean or the proportion exceeding a fixed threshold, using a simple weighted average. For example, the mean of the population of concern can be estimated as

$$\bar{x}_{ATP} = \Sigma_g \; P_g \, \bar{x}_g,$$

where $\Sigma_g$ represents summation over the population of concern groups indexed by *g*, $P_g$ is the proportion of the population of concern that belongs to group *g*, and $\bar{x}_g$ is the sample mean for group *g*.

Unfortunately, if one is interested in estimating a non-linear statistic for the population of concern, such as the variance or a percentile, this technique is not algebraically correct. However, lacking any other information from the sample, calculating this type of weighted average to estimate a non-linear population of concern characteristic is better than making no adjustment at all for known population differences. In the case of the population variance, we recommend calculating the weighted average of the group standard deviations, rather than their variances, and then squaring the estimated population of concern standard deviation to get the estimated population of concern variance.

**Raw Data Available.** If one has access to the raw data, not just summary statistics, options for standardization are more numerous and can be made more rigorously. The options depend, in part, on whether or not the data already have statistical analysis weights, such as those appropriate for analysis of data from a probability-based sample survey.

Suppose that one has access to the raw data from a census or from a sample in which all units can be regarded as having been selected with equal probabilities (e.g., a simple random sample). In this case, if one knows the number, $N_g$, of population of concern members in group $g$, then the statistical analysis weight to associate with the $i$-th member of the $g$-th group is

$$W_g\,(i) \;=\; \frac{N_g}{n_g},$$

where the sample contains $n_g$ members of group $g$. Alternatively, if one knows only the proportion of the population and sample that belong to each group, one can calculate the weights as

$$W_g\,(i) \;=\; \frac{P_g}{p_g},$$

where $p_g$ is the proportion of the sample in group $g$. The latter weights differ from those above only by a constant, the reciprocal of the sampling fraction, and will produce equivalent results for means and proportions. However, the former weights must be used to estimate population totals. In either case, the population of concern mean can be estimated as

$$\bar{x}_{ATP} \;=\; \frac{\Sigma_g\, \Sigma_i\;\, W_g(i)\, x_g(i)}{\Sigma_g\, \Sigma_i\;\, W_g(i)},$$

where $x_g(i)$ is the value of the characteristic of interest (e.g., daily tap water consumption) for the $i$-th sample member in group $g$.

In general, one may have access to weighted survey data, such as results from a probability-based sample of the surrogate population. In this case, the survey analysis weight, $w(i)$, for the $i$-th sample member is the reciprocal of that person’s probability of selection with appropriate adjustments to reduce nonreponse bias and other potential sources of bias with respect to the surrogate population. Further adjustments for making inferences to the population of concern are considered below. These results can also be applied to the case of equally weighted survey data, considered above, by considering the survey analysis weight, $w(i)$, to be unity (1.00) for each sample member.

If one knows the distribution of the population of concern with respect to a given characteristic (e.g., the age/race/gender distribution), then one can use the statistical technique of poststratification to adjust the survey data to provide estimates adjusted to that same population distribution (see, e.g., Holt and Smith, 1979).[1] In this case, the weight adjustment factor for each member of poststratum $g$ is calculated as

[1] Sampling variances are computed differently for standardized and poststratified estimates, but these details are suppressed in the present discussion (see, e.g., Shah *et al.*, 1993).

$$A_g = \frac{N_g}{\Sigma_{i \in g} w(i)},$$

where the summation is over all sample members belonging to poststratum *g*. The poststratified analysis weight for the *i*-th sample member belonging to poststratum *g* is then calculated as

$$w_P(i) = A_g \, w(i).$$

Using this weight, instead of the surrogate population weight, *w(i)*, standardizes the survey estimates to the population of concern.

If one knows multiple marginal distributions for the population of concern but not their joint distribution (e.g., marginal age, race, and gender distributions), one can apply a statistical weight adjustment procedure known as raking, or iterative proportional fitting, to standardize the survey weights (see, e.g., Oh and Scheuren, 1983). Raking is an iterative procedure for scaling the survey weights to known marginal totals.

If one knows population of concern subgroup totals for continuous variables, a generalized raking procedure can be used to standardize the survey weights to known distributions of categorical variables as well as known totals for continuous variables. The generalized raking procedures utilize non-linear, exponential modeling (see, e.g., Folsom, 1991 and Deville *et al.*, 1993).

Of course, none of these standardization procedures results in inferences to the population of concern that are as defensible as those from a well-designed sample survey selected from a sampling frame that completely and adequately covers the population of concern.

### 5.2 Adjustments to Account for Time-Unit Differences.

A common way in which the surrogate population and population of concern may differ is in the time unit of (desired) observation. Probably the most common situation occurs when the study data represent short-term measurements but where chronic exposures are of interest. In this case, some type of model is needed to make the time-unit inference (e.g., from the distribution of person-day or person-week exposures to the distribution of annual or lifetime exposures). In general, it is convenient to break down the overall inference into two components: from the time unit of measurement to the time duration of the study (data to the surrogate population), and from the time duration of the surrogate population to the time unit of the population of concern. For specificity, let t denote the observation time (e.g., a day or a week); let $\tau$ denote the duration of the study (i.e., $\tau$ is the time duration associated with the surrogate population); and let T denote the time unit of the population of concern (e.g., a lifetime). In the case of chronic exposure concerns, $t<\tau<T$.

Suppose that N denotes the number of persons in the surrogate population, and assume there are (conceptually) K disjoint time intervals of length t that surrogate population $\tau$ (i.e., Kt=τ). Thus a census of the surrogate population would involve NK short-term measurements (of exposures or of exposure factors). This can be viewed as a two-way array with N rows (persons) and K columns (time periods). Clearly, the distribution of these NK measurements, whose mean is the grand total over the NK cells divided by NK, encompasses both variability among people and variability among time periods within people (and in practice, measurement error also). The average across the columns for a given row (the marginal mean) is the average exposure for the given person over a period of length τ. Since the mean of these τ-period "measurements" over the N rows leads to the same mean as before, it is clear that the mean of the t-time measurements and the mean of the τ-time measurements is the same. However, unless there is no within-person variability, the variability of the longer τ-period measurements will be smaller than the variability of the shorter t-period measurements. If the distribution of the shorter term measurements is right-skewed, as is common, then one would expect the longer term distribution to exhibit less skewness. Note that the degree to which the variability shrinks depends on the relation between the within-person and between-person components of variance, which is related to the temporal correlation. For example, if there is little within-person variability, then people with high (low) values will remain high (low) over time, implying that the autocorrelation is high and that the shrinkage in variability in going from days to years (say) will be minimal. If there is substantial within-person variation, then the autocorrelations will be low and substantial shrinkage in the within-person variance (on the order of a t/τ decrease) will occur.

To make this t-to-τ portion of the inference, we therefore would ideally have a valid probability-based sample of the NK person-periods, and data on the t-period exposures or exposure factors would be available for each of these sampling units. As a part of this study design, we would also want to ensure that at least some of persons have measurements for more than one time period, since models that allow the time extrapolation will need data that, in essence, will support the estimation of within-person components of variability. There are several examples of models of this sort, some of which are described below.

Wallace *et al.* (1994) describe a model, which we refer to as the Duan-Wallace (DW) model, in which data over periods of length t, 2t, 3t, etc. (i.e., over any averaging period of length mt) are all conceptually regarded to be approximated by lognormal distributions, with parameters that depend on a "lifetime" variance component and a short term variance component. While such an assumption is theoretically inconsistent if exact lognormality is required, it may nevertheless serve well as an approximation. The basic notion of the DW method is that, while the mean of the exposures stays constant, the variability decreases as the number of periods averaged together increases. Hence it is assumed that the total variability for a distribution that averages over M periods (M=1,2,...) can be expressed in terms of a long-term component and a short-term component. Let $\gamma_L$ and $\gamma_S$ denote, respectively, the log-scale variances for these two components. Under the lognormal model, Wallace *et al.* show that the log-scale variance for the M-period distribution (i.e., the distribution that averages over M periods) is given by

$$V_M = \gamma_L + \log[1 + \frac{\exp(\gamma_S) - 1}{M}].$$

Note that an implication of the DW model is that the geometric means for the various distributions will increase as M increases. In fact, the geometric mean (gm) associated with the average of M short-term measurements will be

$$gm(M) = \bar{Y} \exp[-V_M /2]$$

where $\bar{Y}$is the overall population mean of the exposures. As a consequence, if data are adequate for estimating the variance components (and the mean of the exposures), then an estimated distribution for any averaging time can be inferred. In particular, the DW method can be applied if data are available for estimating $V_M$ for (at least) two values of M, since one is then able to determine values of the two variance components. For instance, if two observations per person are available, one can estimate population mean and the population log-scale variance ($V_1$) for single measurements (M=1), and by averaging the two short-term measurements and then taking logs, one can estimate the population log-scale variance, $V_2$. (Sampling weights should be used when applicable.). By substituting into the above $V_T$ equation for T=1 and T=2, the following formulas for estimating the variance components can be determined:

$$\hat{\gamma}_S = -\log[2\exp(\hat{V}_2 - \hat{V}_1) - 1]$$

and

$$\hat{\gamma}_L = \hat{V}_1 - \hat{\gamma}_S..$$

The distribution for any averaging time can then be estimated by choosing the appropriate M (e.g., M=365 if the measurement time is one day) and substituting estimates into the $V_M$ equation above. Similarly, a "lifetime" distribution (also assumed to be lognormal) is then estimated by letting M go to infinity (i.e., the influence of the short term component vanishes). Wallace *et al.*(1994) caution that the data collection period should encompass all major long-term trends such as seasonality.

Clayton *et al.* (1998) describe a study of personal exposures to airborne contaminants that employs a more sophisticated study design and model (that requires more data); the goal was to estimate distributions of annual exposures from 3-day exposure measurements collected throughout a 12-month period. Two measurements per person (in different months) were available for some of the study participants. A multivariate lognormal distribution was assumed; the lognormal parameters for each month's data were estimated, along with the correlations for each monthly lag (assumed to depend only on the length of the lag). Simulated data were generated from this multivariate distribution for a large number of "people;" each "person's" exposures were then averaged over the 12 months. This approach assumes that the an average over 12 observations, one per month, produces an adequate approximation to the annual distribution of exposures. The model results were compared to those obtained via a modification of the DW model.

Buck *et al.* (1995, 1997) describe some general models (e.g., lognormally is not assumed); these, too, require multiple observations per person, and if the within-person variance is presumed to vary by person, then a fairly large number of observations per person may be needed. These papers give some insight into how estimated distributional parameters based on the short-term data relate to the long-term parameters. Reports by Carriquiry *et al.* (1995, 1996), Carriquiry (1996), and a paper by Nusser *et al.* (1996) deal with the some of the same issues in the context of estimating distributions of “usual” food intake and nutrition from short-term dietary data.

The second part of the inference — extrapolation from study time period (of duration $\tau$) to the longer time T — is likely to be much less defensible than the first part, if $\tau$ and T are very different. This part of the inference is really an issue of temporal coverage. If the study involves person-day measurements conducted over a two-month period in the summer, and annual or lifetime inferences are desired, then little can be said regarding the relative variability or mean levels of the short-term and T-term data, basically because of uncertainty regarding the stationarity of the exposure factor over seasons and years. The above-described approach of Wallace *et al.*, for instance, includes statements that recognize the need for a population stationarity assumption that essentially requires that the processes underlying the exposure factor data that occur outside the time period of the surrogate population be like those that occur within the surrogate population. Applying some of the above methods on an age-cohort-specific basis, and then combining the results over cohorts, offers one possible way of improving the inference (e.g., see Hartwell *et al.*, 1992).

## 6. SUMMARY AND CONCLUSIONS

Representativeness is concerned with the degree to which “good” inferences can made from a set of exposure factor data to the population of concern. Thus evaluating representativeness of exposure factor data involves achieving an understanding of the source study, making an appraisal of the appropriateness of its internal inferences, assessing how and how much the surrogate population and population of concern differ, and evaluating the importance of the differences. Clearly, this can be an extremely difficult and subjective task. It is, however, very important, and sensitivity analyses should be included in the risk assessment that reflect the uncertainties of the process.

In an attempt to ensure that all aspects of representativeness are considered by analysts, we have partitioned the overall inferential process into components, some of which are concerned with design and measurement features of the source study that affect the internal inferences, and some of which are concerned with the differences between the surrogate population and the population of concern, which affect the external portion of the inference. We also partition the inferential process along the lines of the population characteristics — individual, spacial, and temporal — in an attempt to assess where overlaps and gaps exist between the data and the population of concern. In the individual and spatial characteristics, representativeness involves consideration of bounds and coverage issues. In the temporal characteristic, these same issues (i.e., study duration and currency) are important, but the time unit associated with the measurements or observations is also important, since time unit differences often occur between the data and the population of concern. Checklists are provided to aid in assessing the various components of representativeness.

When some aspect of representativeness is lacking in the available data, assessors are faced with the task of trying to make the data "more representative." We describe several techniques (and cite some others) for accomplishing these types of tasks; generally making such adjustments for known differences will reduce bias. However, it should be emphasized that these adjustment techniques cannot guarantee representativeness in the resultant statistics. For supporting future, large-scale (e.g., regional or national) risk assessments, one of the best avenues for improving the exposure factors data would be to get assessors involved in the design process -- so that appropriate modifications to the survey designs of future source studies can be considered. For example, the design might be altered to provide better coverage of certain segments of the population that may be the focus of risk assessments (e.g., more data on children could be sought). The use of multiple observations per person also could lead to improvement in those assessments concerned with chronic exposures.

## 7. BIBLIOGRAPHY

American Society for Quality Control (1994). *American National Standard: Specifications and Guidelines for Quality Systems for Environmental Data and Environmental Technology Programs (ANS!/ASQC E4)*. Milwaukee, WI.

Barton, M., A. Clayton, K. Johnson, R. Whitmore (1996). "G-5 Representativeness." Research Triangle Institute Report (Project 91U-6342-116), prepared for U.S. EPA under Contract No. 68D40091.

Buck, R.J., K.A. Hammerstrom, and P.B. Ryan (1995). "Estimating Long-Term Exposures from Short-Term Measuremetns." *Journal of Exposure Analysis and Environmental Epidemiology*, Vol. 5, No. 3, pp. 359-373.

Burmaster, D.E. and A.M. Wilson (1996). "An Introduction to Second-Order Random Variables in Human Health Risk Assessments." *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 892-919.

Carriquiry, A.L. (1996). "Assessing the Adequacy of Diets: A Brief Commentary" (Report prepared under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Carriquiry, A.L., J.J. Goyeneche, and W.A. Fuller (1996). "Estimation of Bivariate Usual Intake Distributions" (Report prepared under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Carriquiry, A.L., W.A. Fuller, J.J. Goyeneche, and H.H. Jensen (1995). “Estimated Correlations Among Days for the Combined 1989-91 CSFII” (Dietary Assessment Research Series Report 4 under Cooperative Agreement No. 58-3198-2-006, Agricultural Research Service, USDA, and Iowa State University).

Clayton, C.A., E.D. Pellizzari, C.E. Rodes, R.E. Mason, and L.L. Piper (1998). “Estimating Distributions of Long-Term Particulate Matter and Manganese Exposures for Residents of Toronto, Canada.” Submitted to *Atmospheric Environment*.

Cohen, J.T., M.A. Lampson, and T.S. Bowers (1996). “The Use of Two-Stage Monte Carlo Simulation Techniques to Characterize Variability and Uncertainty in Risk Analysis.” *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 939-971.

Corder, L.S., L. LaVange, M.A. Woodbury, and K.G. Manton (1990). “Longitudinal Weighting and Analysis Issues for Nationally Representative Data Sets.” Proceedings of the *American Statistical Association, Section on Survey Research*, pp. 468-473.

Deville, J., Sarndal, C., and Sautory, O. (1993). “Generalized Raking Procedures in Survey Sampling.” *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 1013-1020).

Ferson, Scott (1996). “What Monte Carlo Methods Cannot Do.” *Human and Ecological Risk Assessment*, Vol. 2, No. 4, pp. 990-1007.

Folsom, R.E. (1991). “Exponential and Logistic Weight Adjustments for Sampling and Nonresurrogate populationonse Error Reduction.” Proceedings of the *Social Statistics Section of the American Statistical Association*, 197-202.

Francis, Marcie and Paul Feder, Battelle Memorial Institute (1997). “Development of Long-Term and Short-Term Inhalation Rate Distributions.” Prepared for Research Triangle Institute.

Hartwell, T.D., C.A. Clayton, and R.W. Whitmore (1992). “Field Studies of Human Exposure to Environmental Contaminants.” Proceedings of the *American Statistical Association, Section on Statistics and the Environment*, pp. 20-29.

Holt, D. and Smith, T.M.F. (1979). “Post Stratification.” *Journal of the Royal Statistical Society*, Vol. 142, Part 1, pp. 33-46.

Kendall, M.G. and W.R. Buckland (1971). A Dictionary of Statistical Terms. Published for the International Statistical Institute, Third Edition, New York: Hafner Publishing Company, Inc., p. 129.

Kruskal, W. and F. Mosteller (1979). “Representative Sampling, I: Non-Scientific Literature.” *International Statistical Review*, Vol. 47, pp. 13-24.

Kruskal, W. and F. Mosteller (1979). “Representative Sampling, II: Scientific Literature, Excluding Statistics.” *International Statistical Review*, Vol. 47, pp. 111-127.

Kruskal, W. and F. Mosteller (1979). “Representative Sampling, III: The Current Statistical Literature.” *International Statistical Review,* Vol. 47, pp. 245-265.

Nusser, S.M., A.L. Carriquiry, K.W. Dodd, and W.A. Fuller (1996). “A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions.” *Journal of the American Statistical Association*, Vol. 91, No. 436, pp. 1440-1449.

Oh, H.L. and Scheuren, F.J. (1983). “Weighting Adjustment for Unit Nonresponse.” In: *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, Madow, W.G., Olkin, I., and Rubin, D.B., eds., Academic Press, New York, NY, pp. 143-184.

Shah, B., R. Folsom, L. LaVange, S. Wheeless, K. Boyle, and R. Williams (1993). *Statistical Methods and Mathematical Algorithms Used in SUDAAN*. Research Triangle Institute, Research Triangle Park, NC.

Smith, F., K. Kulkarni, L.E. Myers, and M.J. Messner (1988). Evaluating and Presenting Quality Assurance Sampling Data. In Keith, L.H. (Ed.), Principles of Environmental Sampling, American Chemical Society, ACS Professional Reference Book, pp. 157-168.

Stanek, III, E.J. (1996). “Estimating Exposure Distributions: A Caution for Monte Carlo Risk Assessment.” *Human and Ecological Assessment*, Vol. 2, No. 4, pp. 874-891.

Wallace, L.A., Naihua Duan, and Robert Ziegenfus (1994). “Can Long-Term Exposure Distributions Be Predicted from Short-Term Measurements?.” *Risk Analysis*, Vol. 14, No. 1, pp. 75-85.

## CHECKLIST I. ASSESSING INTERNAL REPRESENTATIVENESS: POPULATION SAMPLED VS. POPULATION OF CONCERN FOR THE SURROGATE STUDY

- What is the study population?
  - What are the individual characteristics (i.e., defined by demographic, socioeconomic factors, human behavior and other study design factors)?
  - What are the spatial characteristics?
  - What are the temporal characteristics?
  - What are units of observation (e.g., person-days or person-weeks)?
  - What, if any, are the population subgroups for which inferences were especially desired?

- Are valid statistical inferences to the study population possible?
  - Was the whole population sampled (i.e., a census was conducted) used?
  - If not was the sample design appropriate and adequate?
    - Was a probability sample used? If not, how reasonable does the method of sample selection appear to be?
    - Was the response rate satisfactory?
    - Was the sample size adequate for estimating central tendency measures?
    - Was the sample size adequate for estimating other types of parameters (e.g., upper percentiles)?
    - For what population or subpopulation size was the sample size adequate for estimating measures of central tendency?
    - For what population or subpopulation size was the sample size adequate for estimating other types of parameters (e.g., upper percentiles)?
    - What biases are known or suspected as a result of the design or implementation or the study? What is the direction of the bias?

- Does the study appear to have and use a valid measurement protocol?
  - What is the likelihood of Hawthorne effects? What impact might this have on bias or variability?
  - What are other sources of measurement errors (e.g., recall difficulties)? What impact might they have on bias or variability?

- Does the study design allow (model-based) inferences to other time units?
  - What model is most appropriate?
  - What assumptions are inherent to the model?

**CHECKLIST II. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN – INDIVIDUAL CHARACTERISTICS**

- How does the population of concern relate to surrogate study population in terms of the individuals' characteristics?
  - Case 1: Are the individuals in the two populations essentially the same?
  - Case 2: Are the individuals in the population of concern a subset of those in the study population? If so, is there adequate information available to allow for the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate data to include only persons in the population of concern and then treat this case as Case 1.)
  - Case 3: Are the individuals in the surrogate study population a subset of those in the population of concern?
  - Case 4: Are two populations disjoint -- in terms of individual characteristics?

- How important is the difference in the two populations (population of concern and surrogate population) with regard to the individuals' characteristics? To what extent is the difference between the individuals of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- in terms of the individuals' characteristics?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

**CHECKLIST III. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN -- SPATIAL CHARACTERISTICS**

- How does the population of concern relate to surrogate population in the spatial characteristics?
  - Case 1: Do they cover the same geographic area?
  - Case 2: Is the geographic area of the population of concern a subset of the area of surrogate population? If so, is there adequate information available to allow the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate population to include only regions or types of geographic areas in the population of concern and then treat this case as Case 1.)
  - Case 3: Is the geographic area covered by the surrogate population a subset of that covered by the population of concern?
  - Case 4: Are two populations disjoint -- in the spatial characteristics?

- How important is the difference in the two target populations with regard to the spatial characteristics? To what extent is the difference in the spatial characteristics of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- in terms of the spatial characteristics?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

**CHECKLIST IV. ASSESSING EXTERNAL REPRESENTATIVENESS: SURROGATE POPULATION VS. EXPOSURE ASSESSOR'S POPULATION OF CONCERN -- TEMPORAL CHARACTERISTICS**

- How does the population of concern relate to surrogate population in terms of currency and temporal coverage (study duration)?
  - Case 1: Are the duration and currency of the surrogate data compatible with the population of concern needs?
  - Case 2: Is the temporal coverage of the population of concern a subset of the surrogate population? If so, is there adequate information available to allow the analysis of the population of concern? (Note: If so [Case 2a], we can redefine the surrogate population to include only time periods (e.g., seasons) of interest to the assessor and then treat this case as Case 1.)
  - Case 3: Is the temporal coverage of the surrogate population a subset of that covered by the population of concern?
  - Case 4: Are the two populations disjoint — in terms of study duration and currency?

- How does the population of concern relate to surrogate population in terms of the time unit (either the observed time unit or, if appropriate, a modeled time unit)?
  - Case 1: Are the time units compatible?
  - Case 2: Is the time unit for the population of concern shorter than that of the surrogate population? If so, are data available for the shorter time unit associated with the population of concern. (If so [Case 2a], this can be treated as Case 1.)
  - Case 3: Is the time unit for the population of concern longer than that of the surrogate population?

- How important is the difference in the two populations (i.e., population of concern and surrogate population) with regard to the temporal coverage and currency? To what extent is the difference in the temporal coverage and currency of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- to account for differences in temporal coverage or currency?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

- How important is the difference in the two populations (i.e., population of concern and surrogate population) with regard to the time unit of observation? To what extent is the difference in the observation time unit of the two populations expected to affect the population parameters?
  - With respect to central tendency of the two populations?
  - With respect to the variability of the two populations?
  - With respect to the shape and/or upper percentiles of the two populations?

- Is there a reasonable way of adjusting or extrapolating from the surrogate population to the population of concern -- to account for differences in observation time units?
  - What method(s) should be used?
  - Is there adequate information available to implement it?

# Issue Paper on Empirical Distribution Functions and Non-parametric Simulation

## Introduction

One of the issues facing risk assessors relates to the best use of empirical distribution functions (EDFs) to represent stochastic variability intrinsic to an exposure factor. Generally, one of two situations occurs. In the first situation, the risk assessor is reviewing an assessment in which an EDF has been used. The risk assessor needs to make a judgement whether or not the use of the EDF is appropriate for this particular analysis. In the second situation, the risk assessor is conducting his/her own assessment and must decide whether a parametric representation or non-parametric representation is best suited to the assessment. The objective of this issue paper is to help focus discussion on the key issues and choices facing the assessor under these circumstances.

We make the initial assumption that the data are sufficiently representative of the exposure factor in question. Here, representative is taken to mean that the data were obtained as a simple random sample of the relevant characteristic of the correct population, that the data were measured in the proper scale (time and space), and that the data are of acceptable quality (accuracy and precision).

We also make the assumption that the analysis involves an exposure/risk model which includes additional exposure factors, some of which also exhibit natural variation. Ultimately, we are interested in estimating some key aspects of the variation in predicted exposure/risk. As a minimum, we are interested in statistical measures of central tendency (e.g., median), the mean, and some measure of plausible upper bound or high-end exposure (e.g., 95th, 97.5th, or 99th percentiles of exposure). Thus, how variable factors algebraically and statistically interact is important.

Further, we assume that Monte Carlo methods will be used investigate the variation in exposure/risk. Obviously, other methods can be used, but it is clear from experience that simulation-based techniques will be used in the vast majority of applications.

Conventional wisdom advises that when there is an underlying theory supporting the use of a particular theoretical distribution function (TDF), then the data should be used to fit the distribution and that distribution should be used in the analysis. For example, it has been argued that repeated dilution and mixing of an environmental pollutant should eventually result in a lognormal distribution of concentrations. While this is an agreeable concept in principle, it is rare situation when a theory-based TDFs are available for particular exposure factors. Furthermore, theory-based TDFs are often only valid in the asymptotic sense. Convergence is may be very slow, and, in the early stages, the data may be very poorly modeled by the asymptotic form of the TDF. For this issue paper, we assume that no theory-based TDFs are available.

The issue paper is written in two parts. Part I addresses the strengths and weakness of empirical distribution functions; Part II addresses issues related to judging quality of fit for theoretical distributions.

# Part I. Empirical Distribution Functions

**Definitions**. Given representative data, X = {$x_1$, $x_2$, ···, $x_n$ }, the risk assessor has two basic techniques for representing an exposure factor in a Monte Carlo analysis:

**parametric methods** which attempt to characterize the exposure factor using a TDF. For example, a lognormal, gamma, or Weibull distribution is used to represent the exposure factor, and the data are used to estimate values for its intrinsic parameters.

**non-parametric methods** which use the sample data to define an empirical distribution function (EDF) or modified version of the EDF.

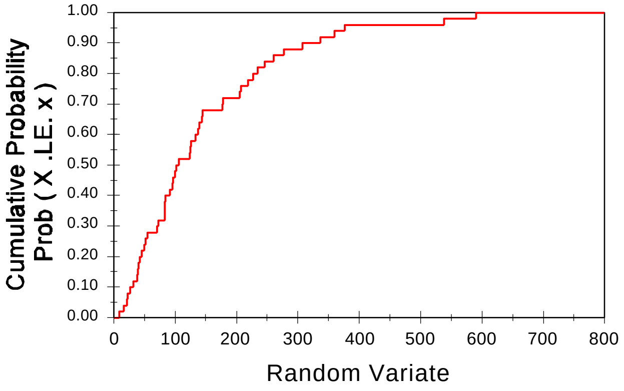
**EDF**. Sorted from smallest to largest, $x_1 \le x_2 \le \cdots x_n$, the EDF is the cumulative distribution function defined by

$$\hat{F}(x) = \frac{number\ \ of\ \ x_k \le x}{n} \qquad or \qquad \hat{F}(x) = \frac{1}{n}\sum_{k=1}^{n} H(x - x_k)$$

where H(u) is the unit step function which jumps from 0 to 1 when u ≥ 0. The values of the EDF are the discrete set of cumulative probabilities (0, 1/$n$, 2/$n$, ···, $n$/$n$). Figure 1 illustrates a basic EDF for 50 samples drawn from lognormal distribution with a geometric mean of 100 and a geometric standard deviation of 3, i.e., X ~ LN(100,3).

In a Monte Carlo simulation, an EDF is generated by randomly sampling the raw data with replacement (simple bootstrapping) so that each observation in the data set, $x_k$, has an equal probability of selection, i.e., $prob(x_k) = 1/n$.

Figure 1. Example of EDF



**Properties of the EDF**. The following summarizes some of the basic properties of the EDF:

1. Values between any two consecutive samples, $x_k$ and $x_{k+1}$ cannot be simulated, nor can values smaller than the sample minimum, $x_1$, or larger than the sample maximum, $x_n$, be generated, i.e., $x \geq x_1$ and $x \leq x_n$

2. The mean of the EDF is equal to the sample mean. The variance of the EDF mean is always smaller than the variance of the sample mean; it is equal to $(n-1)/n$ times the variance of the sample mean.

3. The variance of the EDF is equal to $(n-1)/n$ times the sample variance.

4. Expected values of the EDF percentiles are equal to the sample percentiles.

5. If the underlying distribution is skewed to the right (as are many environmental quantities), the EDF will tend to under-estimate the true mean and variance.

Figures 2 and 3 below illustrate typical Monte Carlo behavior of the EDF in reproducing the sample mean, variance, and 95th percentile of the underlying sample. Here X ~ LN(100,3) with a sample size of $N = 100$ and the relative error is defined as 100× [simulated−sample]/sample. The oscillatory nature of the simulated 95th percentile reflects the normalized magnitude of the difference between adjacent order statistics in the sample, $x_{(95)}$, and $x_{(96)}$ and shows the Monte Carlo estimate flip-flopping between these two ranks

Figure 2. Convergence of the Mean and Variance

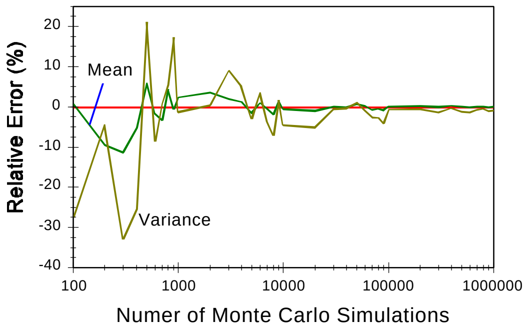
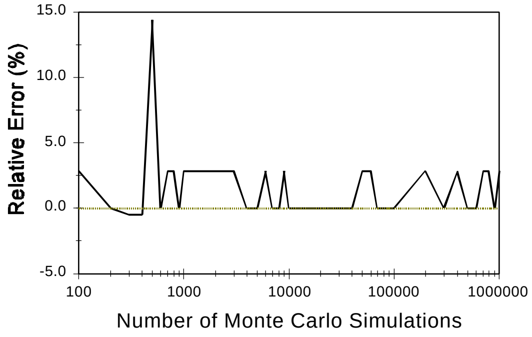


Figure 3. Convergence of the 95th Percentile

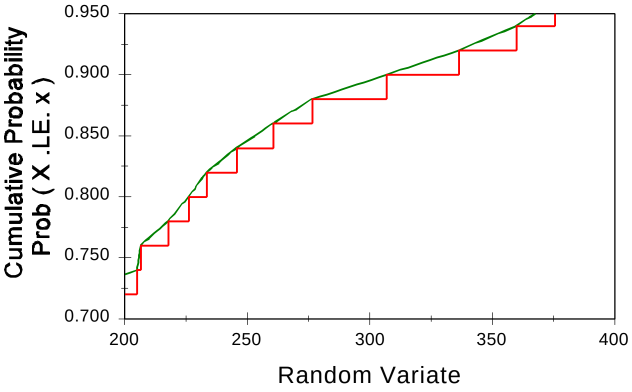


**Linearly Interpolated EDF (Linearized EDF).** For continuous random variables, it may be troubling to define the EDF as a step function and so extrapolation is often used to estimate the probabilities of values in between sample values. Generally, for values between observations, linear interpolation is favored, although higher order interpolation is sometimes used. Figure 4 compares a linearly interpolated EDF with the basic EDF. The linearly interpolated EDF will tend to underestimate the sample mean and variance. It will converge to the appropriate sample percentile, but take longer to do so when compared to the simple EDF. These differences tend to diminish as the sample size increases. Table 1 illustrates differences between the EDF, linearized EDF and best fit TDF for residential room air exchange rates. The EDF statistics are based on a Monte Carlo simulation with 25,000 replications. Clearly the simple EDF is best at reproducing sample moments and sample percentiles.

| | **ACH** | | | **Best Fit** |
|---|---|---|---|---|
| **Statistic** | **Sample** | | **Linearized** | **Weibull** |
| | **N = 90** | **EDF** | **EDF** | **PDF** |
| mean | 0.6822 | 0.6821 | 0.6747 | 0.6782 |
| variance | 0.2387 | 0.2358 | 0.2089 | 0.2479 |
| skewness | 1.4638 | 1.4890 | 1.2426 | 1.2329 |
| kurtosis | 6.6290 | 6.7845 | 5.6966 | 4.9668 |
| 5% | 0.1334 | 0.1320 | 0.1307 | 0.0881 |
| 10% | 0.1839 | 0.1840 | 0.1840 | 0.1452 |
| 50% | 0.6020 | 0.6160 | 0.6032 | 0.5691 |
| 90% | 1.2423 | 1.2390 | 1.2398 | 1.3592 |
| 95% | 1.3556 | 1.3820 | 1.3600 | 1.6450 |

Table 1 Comparison of key summary statistics

Figure 4. Comparison of Basic EDF and Linearly Interpolated EDF



**Extended EDF.** Neither the simple EDF nor the interpolated EDF can produce values beyond the sample minimum or maximum. This may be an unreasonable restriction in many cases. For example, the probability that a previously observed largest value in a sample based on *n* observations will be exceeded in a sample of *N* future observations may be estimated using the relationship $prob = 1 - n/(N + n)$. If the next sample size is the same as the original sample size, there is a 50% likelihood that the new sample will have a largest value greater than the original sample's largest value. Restricting the EDF to the smallest and largest sample values will produce distributional tails that are too short. In order to get around this problem, one may extend the EDF by adding plausible lower and upper bound values to the data. The actual values are usually based on theoretical considerations or on expert judgement. For right skewed data, adding a new minimum and maximum would tend to increase the mean and variance of the EDF. This same sort or rational is used when continuous, unbounded TDFs are truncated at the low and high end to avoid generating unrealistic values during Monte Carlo simulation (e.g., 15 kg adult males, females over 2.5m tall, etc.)

**Mixed Empirical-Exponential Distribution.** An alternative approach to extending the upper tail of an empirical distribution beyond the sample data has been suggested by Bratley *et al*. In their method, an exponential tail is fit to the last five or ten percent of the data. This method is based on extreme value theory and the observation that extreme values for many continuous, unbounded distributions follow an exponential distribution.

## Starting Points

The following table summarizes the results of an informal survey of experts who were asked to contribute their observations and thoughts on the strengths and weaknesses of EDFs by addressing a list of questions and issues. Based on this survey:

1. The World seems to be divided into TDF'ers and EDF'ers.

2. There are no clear-cut, unambiguous statistical reasons for choosing EDFs over TDFs or vice versa.

3. Many of the criticisms leveled at EDFs also apply to TDFs (e.g., the data must be simple random samples)..

4. One aspect of which may have important implications for our discussion is the nature of the decision and how sensitive an outcome is to the choice of an EDF.

5. Generally, contributors did not express much support for either the linearized EDF or the extended EDF. Why they seem to be comfortable with TDFs, which essentially interpolate between data points as well as extrapolated beyond the data, is unclear.

| Issue | Comments |
|---|---|
| 1. EDFs provide complete representation of the data without any loss of information. | Yes, but perhaps an incomplete representation of what is known about the quantity for which the distribution is needed. |
| 2. EDFs do not depend on any assumptions associated with parametric models. | One has to assume a representative random sample.<br>As another example, advantage 2 (EDFs do not depend on parametric assumptions) is true and is a well-known advantage. Less well known is that almost all non-parametric procedures make some strong assumptions. Technically, a parametric situation is one where you limit the class of possible probability distributions to a collection that can be described in a natural way using a finite number of real numbers, or parameters. In common non-parametric situations (such as comparing medians of two sets of data) the data are modeled by pairs of distributions, but there is still a restriction, such as that the members of each pair are the same distribution except for a change of location. Furthermore, using an EDF is something entirely different than the set of assumptions you make about the class of possible distributions. Usually, you use an EDF as a tool to make an estimate: that is, as a computational device. |
| 3. For large samples, EDFs converge to the true distribution for all values of X. | Although for most well-behaved distributions it is the case that the EDF converges in probability to the underlying distribution, convergence often requires unrealistic amounts of data. One important issue in risk assessment is the near universal situation of having too few data. This usually means we are nowhere near a limiting case and that we should beware ALL asymptotic methods, including Maximum Likelihood, without careful evaluation of their applicability to our small data set. EDFs usually converge VERY slowly to the underlying distribution (especially if you're trying to characterize extreme events). Therefore this convergence phenomenon is not very comforting or useful.<br>EDFs are almost useless, except in very large data sets. Accuracy of any interval is driven by a standard deviation of sqrt(n) in that interval. For even a 10% accuracy, with 20 intervals, you would need more than 2,000 underlying observations.<br>This is useless, since "large" is unattainable for all practical purposes, unless you're the Census Bureau. |
| 4. EDFs provide direct information on the shape of the underlying distribution, e.g., skewness and bimodality; EDFs supply robust information on location and dispersion. | Yes, but the confidence limits on those estimates can be quite wide in some cases. For example, a small data set that is negatively skewed could be a random sample from a positively skewed population. |
| 5. An EDF can be an effective indicator of peculiarities (e.g., outliers) | Maybe. Not sure how this is different than when comparing data to a fitted parametric distribution or mixture distributions. |
| 6. An EDF does not involve grouping difficulties and loss of information associated with the use of histograms | True. |

| | |
|---|---|
| 7. Confidence intervals are easily calculated. | For what? how? They can be calculated or simulated for parametric distributions as well. Not sure why this is an advantage for EDFs and not parametric distributions also.<br>It's nice when confidence intervals are easily calculated, but usually the more important criteria are whether they have the coverage claimed of them and how tight the intervals are.<br>Yes, but crude if measurements are limited. The biggest advantage of EDFs you left out: free from subjective model bias. I.e., the choice of parametric form may affect conclusions. |
| 8. EDFs can be sensitive to random occurrences in the data and sole reliance on them can lead to spurious conclusions. This can be especially true if the sample size small | This is true in all cases with small data sets. The best thing is to consider confidence intervals on the distributions to get an idea of whether the occurrences might be random or real.<br>This is ONLY true if the sample size is small. This is the very essence of the issue. |

| | |
|---|---|
| 9. How much data do I need to develop a useful EDF? | What you need is random representative data and to feel comfortable that your data include the lower and upper bounds of the quantity. The number of data points in itself is not particularly important.<br>How many data? Two. This somewhat flippant answer simply highlights the important fact that you need to ask the question in the context of (a) what decision is being made and (b) what its risk function is (how bad is it if the decision is incorrect?). If the risk function is low (it doesn't matter much if we are wrong) and the decision is really obvious, then sometimes all you need is a reality check. Hence the need for one datum. People make mistakes and Murphy's Law applies, so experience dictates a second datum. I know you guys at EPA and in the states are competent and sensible and often very good at this stuff, but there are still many people and many agencies out there that are just too uncomfortable with common sense like this, so it pays to repeat it. (The comment cuts both ways: sometimes I am asked by clients to gather more data to show that they don't have a problem, when all their data point to serious contamination. Most of them back down right away when confronted with the common-sense approach--"you obviously have a problem, so let's talk instead about how to remedy it, since honest statistics won't make it go away.")<br>I would not approach the topic this way. I would ask, instead, how do I characterize an amount of data, and given these summary characteristics, what methods are appropriate.<br>At a minimum 10 points per interval needed, with about 10-20 intervals usually needed for reasonable interpolation of most density curves. For bimodal, etc., double the number of intervals.<br>Gee, that depends. I think the main consideration is the importance of the tails in the decision. If you are going to place a lot of weight on the 99th percentile, then 100 data points are telling you want you want to know. If you are primarily interested in the average or 90th percentile, then 100 data points is pretty good. This is similar to the "how many iterations is enough?" problem. If you have as many data points as iterations, then I think it is pretty hard to justify NOT using an EDF<br>If you are going to place a lot of weight on the 99th percentile, then 100 data points are telling you want you want to know.<br>EXCEPTION: Not with much accuracy. The theory is simple and one example will illustrate the issue. By definition of percentile, there is 0.99 probability that a value above the 99th percentile of the (true) underlying distribution does not occur in a random sample. In a sample of 100 data selected independently from that distribution, values between the 99th and 100th percentiles therefore do not occur with probability (0.99)^100, which is extremely close to 1/e, or almost 40%. Therefore there are almost even odds (2:3) that with 100 data you have not even seen anything as high as the 99th percentile yet. To be fairly sure of seeing a value that high, you need to solve (0.99)^N <= Assurance value (such as 5%) for N. That would require about N=300 points in this example, and even then you only have 95% confidence that you have seen A SINGLE value at or above the 99th percentile. |

| | |
|---|---|
| 10. Should I linearize the EDF between percentiles or use step functions? | A true EDF uses step functions--this is resampling of the data in which each data point has a probability 1/n. The use of linear interpolation will typically lead to lower estimates of the standard deviation, since you are not guaranteed to sample the min and max data points.<br>Now you're going down a slippery slope. As soon as you linearize your EDF you are entering into the land of semi-parametric techniques, smoothing, modeling, and assumptions. You're not using the EDF any more. The EDF is accurately and correctly described by its cumulative distribution function, which will be a step function.<br>If your aren't using a continuous distribution, why not just go with the data? The diversity of distributions is very rich. For example, see Evans, Hasting, and Peacock, Statistical Distributions, 2nd Ed., Wiley (1993) for 39 of them. Using some kind of test for fit of the continuous distribution to your data, e.g., quantiles, you usually can obtain a reasonable fit. See JW Tukey, Exploratory Data Analysis Addison-Wesley (1977). If not, e.g., bimodal, you will have to decompose or transform your data, and you already start to make important assumptions.<br>Smoothing EDFs within the bulk of the probability curve causes no serious errors. Extrapolation beyond the limits of data violates the very concept of EDF, and is intrinsically dependent on the parameterization used.<br>The simple solution is to use the midpoint rule (apply prob. at the interval midpoint). Alternatively, use trapezoidal rule (st. line interpolation). For a continuous curve, a straight line interpolation averages properly and improves discretization bias. I, however, would suggest using resampling as a better approach than smoothing.<br>I usually use percentiles, but you have enough data to use an EDF, then it shouldn't matter much. |
| 11. When the data set is large, should I bin the data into a histogram to speed up the simulation? If so, what defines large? How does it depend on the distribution of the data? | In this case, the difference between step functions and linear interpolations becomes small. Why bin? You lose information that way. If you have large segments of the CDF that are approximately piecewise uniform, then binning the data won't result in much loss of information.<br>here's a lot of literature on binning data, mostly in terms of how the perception of the histogram can change. I would suggest, in the spirit of the response to question 1, that you consider the effect the binning process has on the outcome of your work, since your question really is one of computational practice, not conceptual approach. Bin the data to speed your process (simulation, bootstrapping, whatever) but in a way in which you can demonstrate your answers are not materially different than what you would get with a more accurate procedure. How do you know what a material difference is? Look at your decision space and your risk function.<br>No! This approach causes more mischief in epidemiology than in exposure analysis, but anytime you summarize the data, you lose information. If the data set is large, feel grateful.<br>The intervals or bins used are mathematical estimators of the underlying density or distribution curve. This is a numerical integration or interpolation issue. Typically 10-20 intervals gives good performance on a unimodal density function. Particularly if linear interpolation is used.<br>No. |

| | |
|---|---|
| 12. Should I add a minimum and maximum to the data set so that points outside the observed data can be generated during simulation? The min, max could be based on theoretical considerations or expert judgment. | Why not just use an appropriate parametric distribution instead. This is where the "empirical" approaches fall flat on their face. Some have proposed these bizarre mixed empirical-exponential distributions with exponential and polynomial extrapolations based upon the largest and smallest data points... this can't be defended other than as arbitrary. In contrast, there may be a mechanistic basis for selecting a parametric distribution.<br><br>You're sliding further down the slope. Adding a min or max and using theory or expert judgment seemed to be just what you wanted to avoid by using an EDF. If you're going to do that, you're wide open to criticism. Perhaps better to use some of the other procedures you mention, such as exponential tail fitting. However, if these kinds of procedures will not really change the answer in a material way, go for it.<br><br>Again, no. Let the data talk to you.<br><br>I punt. This is a tail problem that arises when the data really isn't telling you what you want to know. Whatever you base you judgment on will have to be based on other evidence. |
| 13. Should I consider a mixed empirical-exponential distribution? An method for extending the upper tail of an EDF beyond the sample data has been suggested by Bratley et al. In their method, an exponential tail is fit to the last five or ten percent of the data. This method is based on extreme value theory and the observation that extreme values for many continuous, unbounded distributions follow an exponential distribution. The exponential tail is fit so that the mean of the data set is conserved. | I don't like this method as described above. I don't see what it offers in contrast to parametric distributions, and it would seem to open the analyst up for excessive criticism.<br><br>I like the mixed distribution approach (after having carefully read Gnedenko's original paper on extreme value distributions to understand how applicable this approach is). Often you can produce good theoretical and statistical reasons why the tail of your data represents a random sample of extreme values. You need to have this justification, though, since not all probability tails are exponential, and some are very far from it.<br><br>It's no problem to do this, and it may be fun to see what you get, but any conclusions you reach depend entirely on the assumptions in method and your fitting process.<br><br>You could also (and I would somewhat prefer) using more complex (e.g., biphasic) distribution functions that allow more freedom to fit tail data. |
| 14. If I bootstrap and if the exposure variable is continuous, what should I do, if anything, about values in between my data points which will not be simulated in the resampling process? | Probably nothing needs to be done if you assume the data are a representative random sample. The answer will look noisy or jumpy due to the gaps in the data, but that in and of itself is not a bad thing. Use of linear interpolations can lead to different estimates of standard deviations and other statistics when compared with the step-wise EDF.<br><br>You have partially answered this question with the exponential tail fitting suggestion. When you start interpolating and fitting curves to your EDF, you are no longer in the purely parametric realm and you forgo a lot of the EDF advantages you so carefully listed--but sometimes you can't trust using just the EDF.<br><br>As I understand bootstrap, you must generate a distribution, by parameterizing your data, as a first step. This step takes care of interpolation.<br><br>You could bootstrap from percentiles, or take percentiles from bootstraps. I wouldn't think it would make much difference. |

# Part II. Issues Related to Fitting Theoretical Distributions

Suppose the following set of circumstances:

(1) that we have a random sample of an exposure parameter which exhibits natural variation

(2) that the collected data are representative of the exposure parameter of interest (i.e., the data measure the right population, in the right time and spatial scales etc.)

(3) that estimates of measurement error are available.

(4) that there is no available physical model to describe the distribution of the data (i.e., there is no theoretical basis to say that the data are lognormal, gamma, Weibull, etc).

(5) that we wish to characterize and account for the variation in the parameter in an analysis of environmental exposures.

(6) we run the data through our favorite distribution-fitting software and get goodness of fit statistics (e.g., chi-square, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Watson, etc.) and their statistical significance.

(7) rankings based on the goodness of fit results are mixed, depending on the statistic and p-values.

(8) graphical examination of the quality of fit (QQ plots, PP plots, histogram overlays, residual plots, etc) presents a mixed picture, reinforcing the differences observed in the goodness of fit statistics.

## Questions

1). A statistician might say that one should pick the simplest distribution not rejected by the data. But what does that mean when rejection is dependent on the statistic chosen and an arbitrary level of statistical significance?

2). On what basis should it be decided whether or not a data set is adequately represented by a fitted analytic distribution?

3). Specifically, what role should the p-value of the goodness of fit statistic play in that judgment?

4). What role should graphical examination of fit play?

**Respondent #1**
All distributions are, in fact empirical. Parametric distributions are merely theoretical constructs. There is no reason to believe that any given distribution is, in fact, log-normal (or any other specific parametric type). That we agree to call a distribution log-normal is (or at least should be) merely a shorthand by which we mean that it looks sufficiently like a theoretical log-normal distribution to save ourselves the extra work involved in specifying the empirical distribution. Other than analyses where we are dealing strictly with hypothetical constructs (e.g, what if we say that such-and-such distribution is lognormal and such and such distribution is normal....), I can see no theoretical justification for a parametric distribution other than the convenience gained. When the empirical data are sparse in the tails, we, of course, run into trouble in needing to specify an arbitrary maximum and minimum to the empirical distribution. While this may introduce considerable uncertainty, it is not necessarily a more uncertain practice than allowing the parametric construct to dictate the shape of the tails, or for that matter arbitrarily truncating the upper tail of a parametric distribution. This becomes less of a problem if the analysts goal in constructing an input distribution is to describe the existing data with as little extrapolation as necessary rather than to predict the "theoretical" underlying distribution. This distinction gets us close to the frequentist/subjectivist schism where many, if not all MC roads eventually seem to lead.

**Respondent #2**
...if you use p-bounds you don't have to choose a single distribution. You can use the entire equivalence class of distributions (be it a large or small class). I mean, if you can't discriminate between them on the basis of goodness of fit, maybe you do the problem a disservice to try. And operationalizing the criterion for "simplest" distribution is no picnic either.

**Respondent #3**
Why not try the KISS method: Keep It Simple & Sound. The Ranked Order Data assuming uniform probability intervals is a method that makes no assumptions as to the nature of the distribution. I also tends to the true distribution function as the number of data points increases. If you have replicate measurements (on each random sample) then the mean of these should be used.

The method yields simple rapid random number generators and one can obtain and desired statistical parameter of the distribution. However, use of the distribution function in any estimate is advised. Given the high level of approximation and/or bias in most risk assessment data and models, any approximation to the true PDF should be adequate.

There is one occasion when the theoretical PDF may be better than the empirical PDF. That is when it comes from the solution of equations based on fundamental laws constraining the solution to a specified form. Even in this case agreement with data is required. This in not usually the case in risk assessment PDFs.

**Respondent #4**
Since I am blessed not to be a statistician, I have no problem disputing their "statement" about the "simplest" distribution. I don't know what they mean either. What really matters physically is picking a distribution that has the fewest variables and that is easy to apply, given the kind of analysis you want to do. You want one that does not make assumptions in its construction that contradict processes operating in your data. If your are generating equally bad fits with a variety of the usual distributions anyway, by all means chose the one that is easiest to use. For time sliced exposure data, the "right" distribution almost always means a lognormal distribution. A physical basis for the lognormal does exist for exposure data, and empirically, most exposure data fit lognormals. [Your assumption "A" does not hold for typical exposure processes.] Wayne Ott, who probably does not even remember it, taught me this one afternoon in the back of a meeting room. See "A Probabilistic methodology for analyzing water quality effects of urban runoff on rivers and streams," Office of Water, February 15, 1984. Just tell people that you have used a lognormal distribution for convenience, although it does not fit particularly well, then provide some summary statistics that describe the poorness of fit.

Problems begin when you get a poor fit to a lognormal distribution but a good fit with a different distribution. Say you get a better fit to the Cauchy distribution, because the tails of your pdf have more density. Now things get more fun. Statisticians would say that you should use the Cauchy distribution, because it is a better fit. I say that you should still use the lognormal, because you can interpret manipulations of the data more easily, and just note that the lognormal fit is poor. Problems will arise, however, if you want to reach conclusions that rely on the tails of the distribution, and you use the lognormal pdf formulation, instead of your actual data. I somewhat anticipated your dilemma in my previous E-mail to you. If you don't need to use a continuous distribution, just go with the data!"

For time dependent exposure data, the situation gets much more complex. I prefer to work with Weibull distributions, but I see lots of studies that use Box-Jenkins models.

And you also asked: On what basis do I decide whether my data are adequately represented by a fitted analytic distribution? Specifically, what role should the p-value of the goodness of fit statistic play in my choice? What role should graphical examination of fit play?

To me, the data are adequately represented, when the analytical distribution adequately fills the role you intend it to have. In other words, if you substitute a lognormal distribution for your data, as a surrogate, then carry out some operations and obtain a result, the lognormal is adequate, unless it leads to a different conclusion than the actual data would support. The same statement is true of any continuous distribution.

Similarly, as a Bayesian, I think that the proper role of a p-value is the role you believe it should play. I don't think that p-values have much meaning in these kinds of analyses, but if you think they should, you should state the desired value before beginning to analyze the data, and not proceed until you obtain this degree of fittedness or better. If small differences in p-value make much difference in your analysis, your conclusions are probably too evanescent to have much usefulness. The quantiles approach that I previously commended to you, is a graphical method. [See J.W. Tukey, Exploratory Data Analysis. Addison-Wesley (1977)]. In it, you would display the distribution of your data, mapped against the prediction from the continuous distribution you have chosen, with both displayed as order statistics. If your data fit your distribution well, the points (data quantiles versus distribution quantiles, will fall along a straight (x=y) line. Systematic differences in location, spread, and/or shape will show up fairly dramatically. Such visual inspection is much more informative than perusing summary statistics. No "statistical fitting" is involved. [Also see J.M. Chambers et al., Graphical Methods for Data Analysis. Cole Publishing (1983)].

**Respondent #5**

I have several thoughts on the goodness of fit question. First, visual examination of the data is likely to yield more insight into the REASONS for the mixed behavior of the various statistics; i.e., in what regions of the variable of interest does a particular theoretical distribution not fit well, and in what direction is the error? Then choosing a particular parametric distribution can be influenced by the purpose of the analysis. For example, if you are interested in tail probabilities, then fitting well in the tails will be more important than fitting well in the central region of the distribution, and vice versa.

A good understanding of the theoretical properties of the various distributions is also handy. For example, the heavy tails of the lognormal mean that the moments can be very strongly influenced by relatively low-probability tails. If that seems appropriate fine; if not the analyst should be aware of that, etc. I don't think there is a simple answer; it all depends on what you are trying to do and why!

**Respondent #6**

In broad overview, I have these suggestions -- all of which are subject to modification, depending on the situation.

1. Professional judgment is **unavoidable** and is **always** a major part of every statistical analysis and/or risk assessment. Even a (dumb) decision to rely **exclusively** on one particular GOF statistic is an act of professional judgment. There is no way to make any decision based exclusively on "objective information" because the decision on what is considered objective contains unavoidable subjective components. There is no way out of any problem except to use and to celebrate professional judgment. As a profession, we risk assessors need to get over this hang up and move ahead.

2. It is **always** necessary and appropriate to fit several different parametric distributions to a data set. We make choices on the adequacy of a fit by comparison to alternatives. Sometimes we decide that one 2-parameter distribution fits well enough (and better than the reasonable

alternatives) so that we will use this distribution. Sometimes we decide that it is necessary to use a more complicated parametric distribution (e.g., a 5-parameter "mixture" distribution) to fit the data well (and better than the reasonable alternatives). And sometimes, we decide that no parametric distribution can do the job adequately well, hence the need for bootstrapping and other methods.

3. The human eye is far, far better at **judging** the overall match (or lack thereof) between a fitted distribution and the data under analysis than any statistical test ever devised. GOF tests are "blind" to the data! We need to visualize, visualize, and visualize the data -- as compared to the alternative fitted distributions -- to **see** how the various fits compare to the data. Mosteller, Tukey, and Cleveland, three of the most distinguished statisticians of the last 50 years, have all stressed the **essential** nature of visualization and human judgment relying thereon (in lieu of GOF tests). BTW, these graphs and visualizations *must* be published for all to see and understand.

4. In situations where no single parametric distribution provides an **adequate** fit to the data, there are several possible approaches to keep moving ahead. Here are my favorites.

A. (standard approach) Fit a "mixture" distribution to the data.

B. Use the two or three or four parametric distributions that offer the most appealing fit in a sensitivity analysis to see if the differences among the candidate distributions really make a difference in the decision at hand. Get the computer to simulate the results of choosing among the different candidate distributions. This leads to keen insights as to the "value of information".

C. (see references below, and references cited therein) By extension of the previous idea, analysts can fit and use "second-order" distributions that contain both **Variability** and **Uncertainty**. These second-order distributions have many appealing properties, especially the property that they allow the analyst to propagate Variability and Uncertainty **separately** so the risk assessor, the risk manager, and the public can all see how the Var and Unc combine throughout the computation / simulation into the final answer.

**Respondent #7**
[RE comments #1, #3, respondent #6]. ... the motivation behind having standardized methods: Professional judgment does not always produce the same result. Your professional judgment does not necessarily coincide with someone else's professional judgment. Surely, you've noticed this. The problem isn't that no one is celebrating their professional judgement - the problem is that we have more than one party.

The bigger and more unique the problem, the less standardization matters. But if you are trying to compare, say, the risk from thousands of superfund sites, you can't very well reinvent risk analysis for every one and expect to get comparable results - whatever you do for one you must do for all.

Have you tried to produce a GOF statistic that matches your visual preference?  I have.  For instance, I think fitting predicted percentiles produces better looking fits than fitting observed values (e.g., maximum likelihood) - because this naturally gives deviations at extreme values less weight - where 'extreme value' is model dependent.